NOVA IMS Information Management School

MAA

Mestrado em Métodos Analíticos Avançados
Master Program in Advanced Analytics

A framework for the Comparative analysis of text summarization techniques

Trijit Ghosh

Dissertation presented as partial requirement for obtaining the master's degree in Data Science and Advanced Analytics

NOVA Information Management School
Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

A FRAMEWORK FOR THE COMPARATIVE ANALYSIS OF TEXT SUMMARIZATION TECHNIQUES

by

Trijit Ghosh
(M20170009)

Dissertation presented as partial requirement for obtaining the master's degree in Data Science and Advanced Analytics

Advisor / Co Advisor: Professor Ricardo Rei; Professor Roberto Henriques

July 2021

ACKNOWLEDGEMENTS

I would first like to thank my thesis advisor Professor Doctor Ricardo Rei of the NOVA Information Management School at Universidade NOVA de Lisboa as he was the one who challenged me to a theme as innovative as the one that gave motto to this master's thesis on text summarization. I want to thank him for encouraging me and motivating me even when the time to devote to this master's thesis was not what was wanted and expected.

Finally, a big thank you to my parents for encouraging me and giving me the chance to do this master's in areas as interesting and exciting as advanced analytics are. That gave me the possibility to have a career that I have dreamed of. It would not be possible without their support.

Contents

List of Tables

List of figures

1. INTRODUCTION

1.1. BACKGROUND

We see that with the boom of information technology and IOT (Internet of things), the size of information which is basically data is increasing at an alarming rate. This information can always be harnessed and if channeled into the right direction, we can always find meaningful information. But the problem is this data is not always numerical and there would be problems where the data would be completely textual, and some meaning has to be derived from it. If one would have to go through these texts manually, it would take hours or even days to get a concise and meaningful information out of the text. This is where a need for an automatic summarizer arises easing manual intervention, reducing time and cost but at the same time retaining the key information held by these texts. In the recent years, new methods and approaches have been developed which would help us to do so. These approaches are implemented in lot of domains, for example, Search engines provide snippets as document previews, while news websites produce shortened descriptions of news subjects, usually as headlines, to make surfing easier.

Broadly speaking, there are mainly two ways of text summarization – extractive and abstractive summarization. Extractive summarization is the approach in which important sections of the whole text are filtered out to form the condensed form of the text. While the abstractive summarization is the approach in which the text as a whole is interpreted and examined and after discerning the meaning of the text, sentences are generated by the model itself describing the important points in a concise way.

1.2. MOTIVATION

As the Internet has grown in popularity, a vast amount of information has become available. Summarizing vast amounts of text is challenging for humans. In this age of information overload, automatic summarizing technologies are in high demand. We will try to focus on various extraction methodologies for single and multi-document summarization in this thesis. Some of the most often used methods, such as topic representation approaches, frequency-driven methods, graph-based and machine learning techniques, will be described. Even though it is difficult to thoroughly explain all of the many algorithms and approaches in this thesis, we will try to give a good overview of recent trends and breakthroughs in automatic summarizing methods, as well as discuss the state-of-the-art and compare various ways.

1.3. OBJECTIVE

The objective of this paper is to provide comparative analysis of different techniques of text summarization used in different scenarios, methodology - to define a set of analysis parameter that can allow us to classify different techniques e.g., complexity, accuracy and speed.

2. EXTRACTIVE SUMMARIZATION

Extractive summarization, as mentioned above, chooses pertinent subsets from the text given, based on some metrics and is thus combined to a condensed form. To understand how summarization systems work, we describe three, fairly, independent tasks which all summarizers perform:

1) Build a transitional depiction of the input text which communicates the most important aspects of the text.
2) Score the sentences supporting the representation.
3) Choose a summary consisting of variety of texts.

2.1. INTERMEDIATE REPRESENTATION

All these summarization techniques will develop some intermediate representations of the given input text supported certain metrics and discern the important sentences supported those metrics. There are two forms of approaches supported the representation: topic representation and indicator representation.

Topic representation methods remodel the text into a transitional characterization and examine the topic(s) given within the text. This method differs in terms of formulation and thereby, complexity, and are divided into frequency-driven approaches, topic word approaches, latent semantic analysis and Bayesian topic models. We take a deeper look into topic representation approaches within the following sections.

Indicator representation approaches describe every sentence as a listing of features (indicators) of importance like sentence length, position within the document, having certain phrases, etc.

2.2. SENTENCE SCORE

When the input text is transformed into a form which the model interprets, a score is assigned to every sentence based on that metric. This score is just a representation of how important a sentence is. These indicators (metric) are derived from mathematical or machine learning models. Once the score of every of the sentences are obtained, they are, then, aggregated or fed into a function which ranks these sentences based on the scores obtained. they're also referred to as indicator weights.

2.3. SUMMARY SENTENCES SELECTION

After the sentences are scored and ranked supported the various metrics or indicators, the important sentences are filtered out. These processes can use different algorithms to separate the sentences supported the ranks and a few other scores as an example redundancy score. Some methods use greedy algorithms to search out the simplest sentences representing the essence of the text. This method will not always be the most effective approach which ignores

certain other indicators. As such, this scenario might be treated as an issue to be fed to a optimization model also which might find important sentences with the optimum values of rank and redundancy scores and other indicators. For instance, context within which the summary is formed can be helpful to choose the important sentences (e.g., newspaper article, email, scientific paper) is another factor which can impact selecting the sentence.

3. TOPIC REPRESENTATION APPROACHES

In this section we describe a few of the foremost widely used topic representation approaches.

3.1. TOPIC WORDS

One of the earliest works in this field was done by (Luhn, 1958). His work mainly focused on bringing the abstract idea of the article. Although, he says its abstract, it's more of an extractive approach. To do this, his initial thought process was to filter out certain words which would best describe the topic. For filtering out those words, a measure was to be introduced and the words which would pass the threshold significance level would be the right candidates for the summary text. The significance of the words was the scoring of words which would tell us how important or significant a word is. The modus operandi of this approach was to first exclude the most frequently occurring words like determiners, prepositions etc. as well the words which were rarely used. After this, the words that would be left would be fed to a function which measures the significance of it by computing the number of times it has occurred in the article. Also, this approach would compute the significance of the sentence based on the positioning of the significant words in the sentence and thus derive the significance of the sentence. Post which, these significant sentences would be used for summarizing the document (Luhn, 1958).

A more advanced version of Luhn method is applying log likelihood test to discern how important a word is. If a word passes this statistical test, then it would be a favorable candidate to be used for summarizing the text and would be called "topic signature" word. This approach has been seen to increase the performances of the model as compared to its earlier version. So, a sentence can be selected in two ways.

1. If the frequency of the topic signatures is more in a sentence
2. If the percentage of topic signatures is high in a given text.

In the first method, the probability of selecting a sentence would be high if the sentence is long and thereby, the count of topic signature words can be high as well while in the second method, the sentence would only be selected if the density of the topic signature is high (A. & K., 2012).

3.2. FREQUENCY-DRIVEN APPROACHES

It is interesting to see that the above method topic words representation assigns binary weights to the words while discerning its importance in the sentence. According to the research, this method has been very efficient as compared to some other methods which we will be discussing below. The methods described below dwell on the concept of assigning real continuous weights to the words. These methods namely word probability and TDIDF are as follows:

3.2.1 Word Probability. In this method, the frequency of the words occurring in a sentence is used as an indicator of the importance of the word. It goes without saying that this number is not binary and is a real continuous number. This method is named as the word probability because in this approach, the probability of the words is calculated as number of times it has occurred in the input divided by the total number of words in the input (the input could be a single document or a multi-document).

$$P(w) = f(w)/N \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Vanderwende proposed the SumBasic system in which the probability of the sentences is going to be calculated as the average probability of the content words in the sentence.

$$g(S_j) = \sum w_i \in S_j \, P(w_i)/|\{w_i|w_i \in S_j\}| \quad \ldots\ldots\ldots\ldots\ldots..(2)$$

This value is assigned as an indicating factor of how important a sentence is. This approach is also greedy and as such, the sentence with the highest probability is selected and their probabilities are updated by squaring their initial probabilities. This makes sure that once a sentence is selected, the chance of selecting it again is low.

$$p_{new}(w_i)=p_{old}(w_i)p_{old}(w_i) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (3)$$

This updated word weight also tells us that the probability of the word selected for the summary is lower than the word occurring once. Also, for the summary, a desired length is taken as a parameter. And this length is then fulfilled by repeating the above steps. This approach of sentence selection used by SumBasic, is a greedy algorithm. Yih came up with an optimization algorithm (to select the sentence) which maximizes the existence of the main words holistically over the whole summary (A. & K., 2012).

3.2.2 TFIDF. TFIDF is a statistical measure which appraises the pertinence of a word in each document among the collection of documents. As stated in (Jones, 2004), the words which are frequently occurring in the corpus is as important as the words whose counts are infrequent as well. This is where we have the terminology ‘specificity’ introduced in this method. As such, both frequent and infrequent words are co-related and used in weighting a word based on this measure. The figures below taken from (Jones, 2004), is a proof of the fact that they are important.

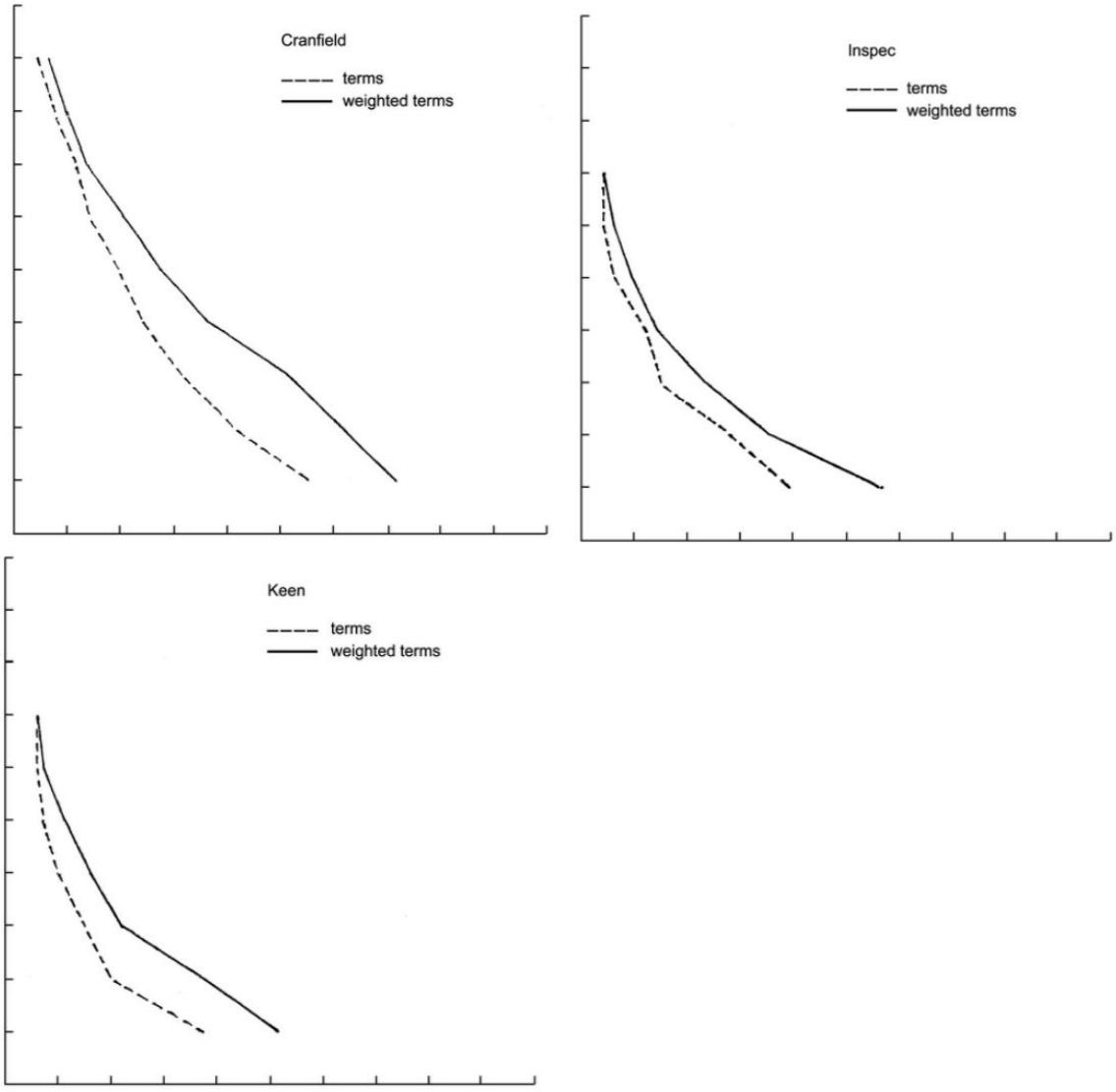


Figure 1 : Weighted Terms v/s Specificity

In the above three figures, it shows that the weighted terms when incorporated with specificity showed increase in performance with different inputs namely Keen, Inspec and Cranfield.

Also, in (Salton & Buckley, 1988), they come with the formulation of specificity, where again it is said that the method was useful when a greater degree of discrimination in the term specificity was incorporated.

Thus in (A. & K., 2012), Term Frequency Inverse Document Frequency method evaluates the significance of words and points out quite repetitive words within the document(s) by assigning low weights to words existing in most documents. the load of every word w in document d is computed as follows:

$$T F * IDF_w = c(w). \log D / d(w)$$

Centroid-based summarization, another, is based on TFIDF topic representation. In this approach, a threshold is defined which comes from experience or rather observation. All the

words whose TFIDF is below that threshold is assigned a weight of zero and is thus ignored. This method is like the topic representation where unimportant words are not considered as well. At the same time, it has similarity with word probability where the word above the threshold value varies between the threshold and one. The sentence scoring function in this approach is given below which is the sum of the weights of the words in a sentence.

$$c_j = \sum d \in C_j d / |C_j| \ldots\ldots\ldots\ldots\ldots\ldots (5)$$

3.3. Latent Semantic Analysis

In (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), they came up with a vectorial way of representing the terms in the corpus with the help of a matrix. This is a representation of term by document matrix which is then further decomposed into ca 100 factors and is used to discern the important sentences. In this paper, they explored lot of mathematical models and interpretation of the term document matrix and were successful in establishing how potent tool LSA can be and recommend using it for retrieval of sentences. Even in (Gong & Liu, 2001), they tried two methods namely relevance measure and latent semantic analysis (LSA). They found that both the methods were equally important in deriving the correct sentences to be used for summarizing a document. The LSA approach first creates a n by m term-sentence matrix, in which each row represents a word from the input (n words) and each column represents a phrase (m sentences). The importance of the word i in sentence j is represented by each entry a_{ij} in the matrix. The TFIDF technique is used to determine the significance of words, and if a sentence does not contain that particular word, the value of that word in the phrase is zero. Then singular value decomposition (SVD) is used on the matrix and converts the matrix A into three matrices: $A=U\Sigma V^T$.

Matrix U(n×m) represents a term-topic matrix having importance of words. Matrix Σ is a diagonal matrix (m×m) where each row i represents the importance of a topic i. Matrix VT is the topic sentence matrix. The matrix D = ΣVT describes how much a sentence represent a topic, thus, d_{ij} shows the importance of the topic i in sentence j. According to (A. & K., 2012), the approach in (Gong & Liu, 2001) has a flaw because a topic may require more than one sentence to carry its information. And as such, other approaches were suggested to ameliorate the performance of LSA-based approaches. One improvisation was to make use of the importance of each topic to evaluate the relative size of the summary that represents the overall topic, which makes it possible to have a varying number of sentences.

Let g be the "weight" function, then $g(s_i)=\sqrt{\sum d^2_{ij}} \ldots\ldots\ldots\ldots\ldots\ldots\ldots (6)$

3.4. Bayesian Topic Models

According to (A. & K., 2012), most of the existing multi-document summarization approaches mainly have 2 disadvantages:

1. They regard the sentences as not related to each other, so topics rooted in the documents are not considered.
2. Sentence scores calculated by most relevant techniques usually do not have very clear probabilistic explanations, and most of them are computed using heuristics.

Bayesian topic models are heuristic in nature and are very efficient and powerful when it comes to representing the topic of the document as a summary. The best part about this method is that they retain the information of the documents which are often missed in other approaches. This benefit of explaining and exhibiting topics with facts allows the development of summarizer systems which, then, discovers the similarities and differences between documents to be applied in summarization [12]. Topic models often use a very different metric for scoring the sentence called Kullbak-Liebler (KL). The KL mainly measures the difference or divergence between the two distributions P and Q [13]. In summarization where we have got probability of words, the KL divergence of Q from P over the words w is defined as:

$$DK\ L(P||Q) = \sum wP(w)\log P(w)/Q(w) \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (7)$$

Where P(w) and Q(w) are probabilities of w in P and Q. KL divergence is a noteworthy technique for grading sentences in the summarization, because it exhibits the idea that good summaries exhibit the same meaning as the original document. It shows that the pertinence and selection of words are impacted when there is a change in the input, i.e. the KL divergence of a good summary and the input will be low.

3.5. BERT

BERT stands for Bidirectional Encoder Representations from Transformers. The BERT architecture is built on top of a transformer of encoders. The flow of the input data is followed according to the given diagram below.

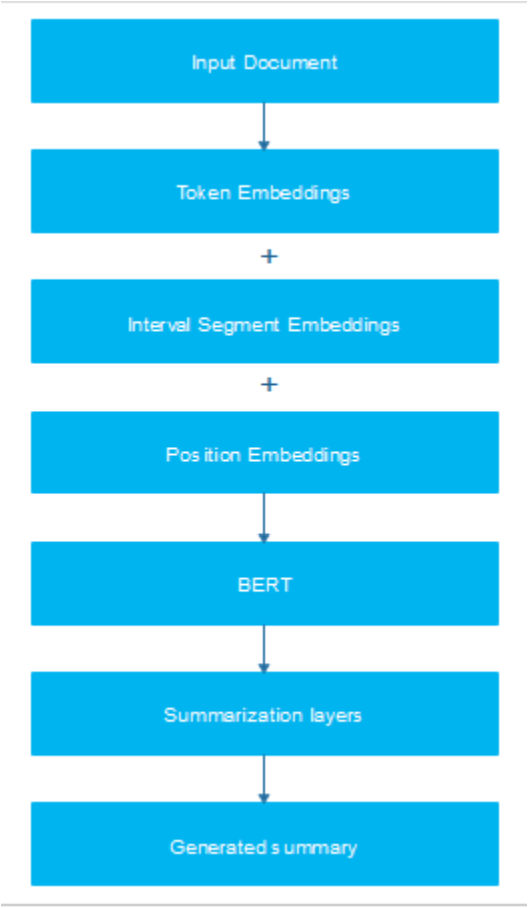
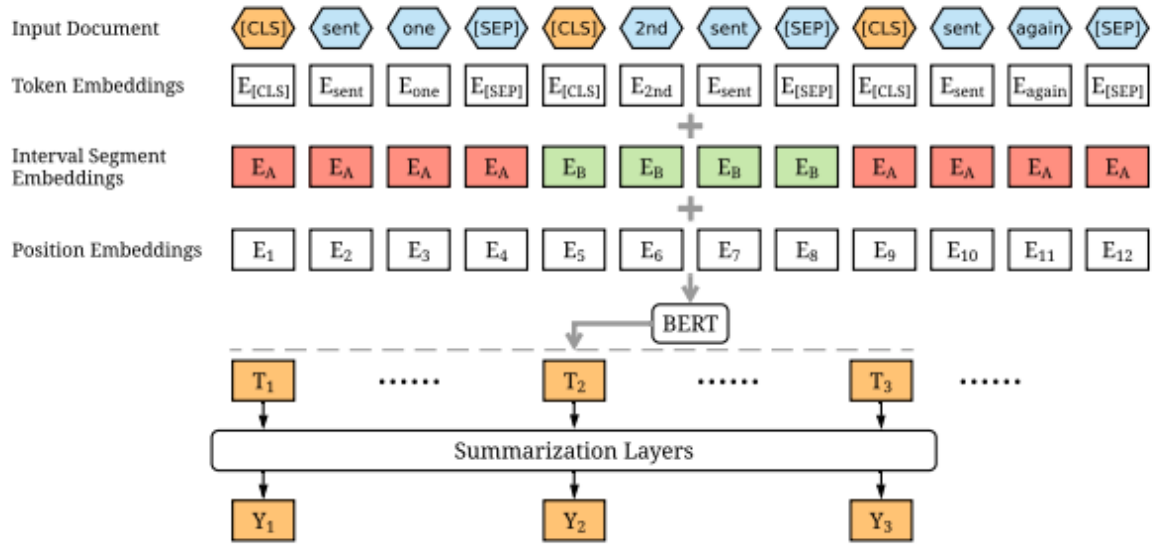


Figure 2 : Bert Embeddings

The input text is first fed into the text processing embeddings namely: -

- *Position Embeddings*
- *Segment Embeddings*
- *Token Embeddings*



The overview architecture of the BERTSUM model.

Figure 3 : Architecture of BERT

The output of these embedding is then fed to the BERT layer which consists of transformers.

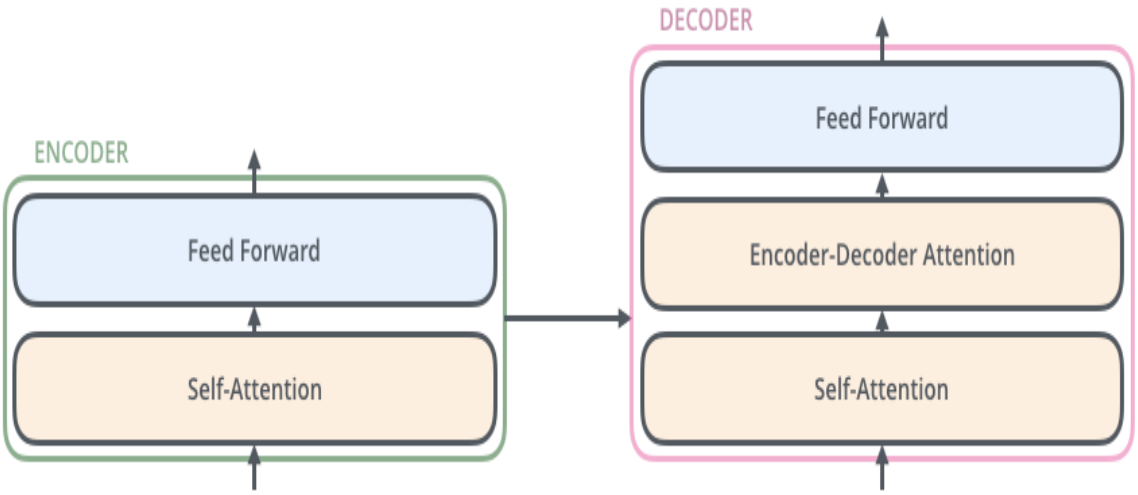


Figure 4 : Encoders and Decoders

A transformer can consist of 12/24 blocks of encoders with 12/16 attention heads and 110/340 million parameters, namely $BERT_{base}$/$BERT_{large}$ respectively. If looked closely, each transformer can be a set of encoders and decoders as shown above in the diagram. It is here that the output of the transformer is fed to the summarization layer.

Pre-training has been quite crucial when it came to language models. There have been

applications of these models such as natural language inference and paraphrasing.
This paper (Devlin, Chang, Lee, & Toutanova, 2018), mainly talks about fine tuning approach with the proposal of BERT as described earlier. The reason it is called bidirectional is because unlike other models where the text is read from left to right or from right to left, the BERT approach reads the sentences in both the directions and tries to understand the context of the words both to its left and right.

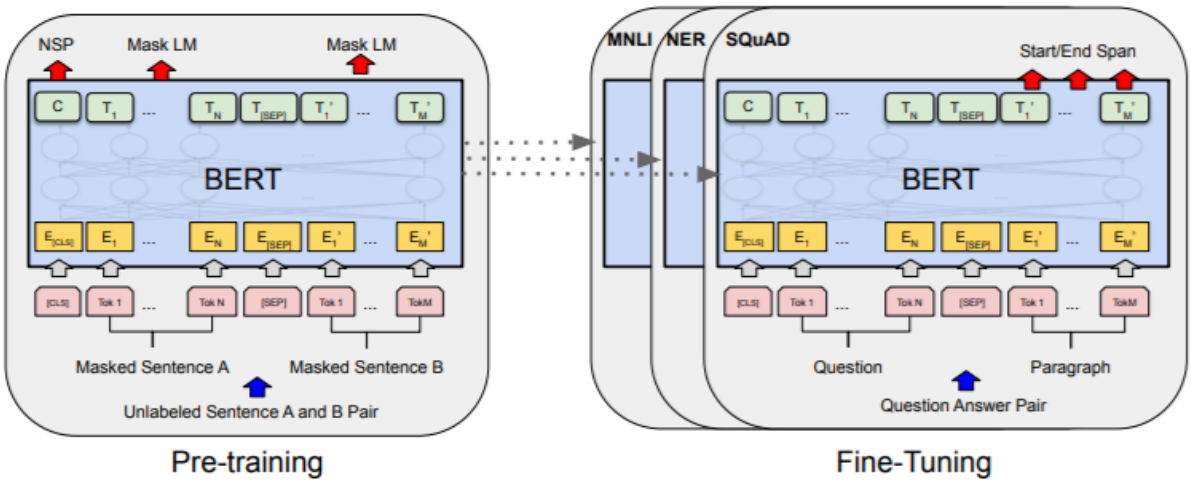


Figure 5 : Overall pre-training and fine-tuning procedures for BERT

In the above diagram, the architecture of the BERT is shown where apart from the output layers, both use the same architecture.

System	MNLI-(m/mm) 392k	QQP 363k	QNLI 108k	SST-2 67k	CoLA 8.5k	STS-B 5.7k	MRPC 3.5k	RTE 2.5k	**Average** -
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Figure 6 : Fine Tuning phase

Source: (Devlin, Chang, Lee, & Toutanova, 2018)

The above figure shows that by working on the fine-tuning phase of the model, the model was able to achieve an accuracy by an overwhelming margin of 4.5% and 7% prior to its state of the art.

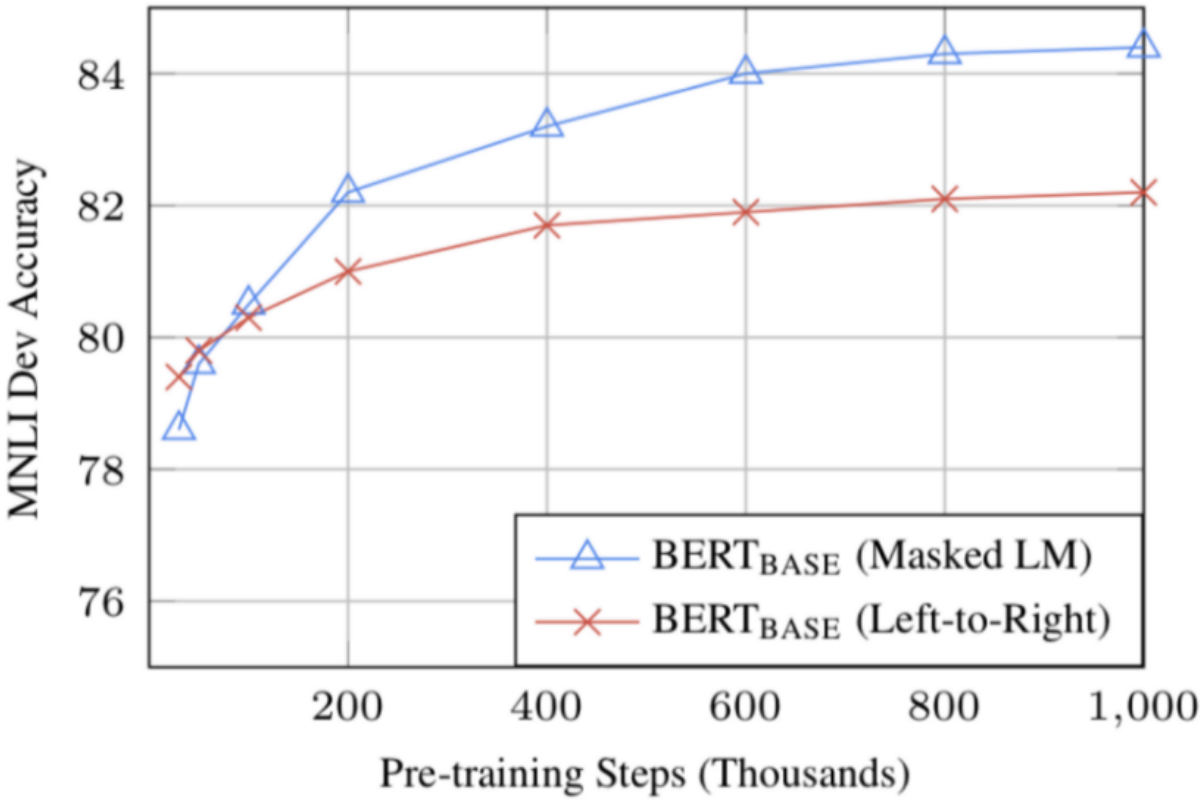


Figure 7 : Accuracy of $BERT_{base}$ on Masked LM and Left-to-Right

Source: (Devlin, Chang, Lee, & Toutanova, 2018)

Although it is to be mentioned that this bi-directional approach takes longer than unidirectional approach which is what has been shown in the above diagram where the BERT MLM method converges slower than left to right but accuracy wise, it is well ahead of the other approach.

This bi-directional approach is, definitely, useful although a bit slow and makes the application of BERT to a wider range of fields.

4. THE IMPACT OF CONTEXT IN SUMMARIZATION

It is quite evident that a context is very useful when it comes to understanding the information presented in a text. In the same way, if models can be empowered with the information of context, then a summarizer system would be able to prune the correct information to summarize the text. For example, according to (Allahyari, et al., 2017), when summarizing blogs, the debates or comments that follow the blog post are useful resources for determining which portions of the blog are critical and intriguing. There is a significant quantity of information available in scientific paper summaries, such as published articles and conference information, that can be used to highlight essential sentences in the original work.

4.1. WEB SUMMARIZATION

If we look at the web pages, we will see that they have lot of objects which is not always possible to summarize for example, pictures, gifs, and some unwanted materials like advertisements which is so not relevant to the information given in the pages. For those kinds of situations, it will be helpful to use the links which direct us to the page to be summarized. These links would give the model a knowledge of the context and would be helpful to provide improved summary for the page. In (Amitay & Paris, 2000), where they talk about website summurization for the first time, they came up with a concept called "THE INCOMMONSENSE SYSTEM". This model incorporates a web crawling system which crawls all the links that link back to the current page and this is how the context is derived. Since then, a lot of different algorithms have been developed which was based on the above principle but has beeen improved.

4.2. SCIENTIFIC ARTICLES SUMMARIZATION

In (Mei & Zhai, 2008), a summurization problem related to scientific papers is studied and presented. Needless, to say with each passing year, new discoveries are being made and new research papers are being published every year. So with this, the daunting task of briefing a scientific papers becomes really challenging. Specially when it comes to including a context in the summary which references a multitude of different research papers. In order to solve this problem, they came with a concept of *impact based summurization* as mentioned in (Mei & Zhai, 2008). This method leverages on the importance of sentence score which has been used in the original paper using the KL divergence method (i.e., finding the similarity between a sentence and the language model). The conclusions made in this paper are substantial porving that this method was useful and could be used for future breifing models for scientific papers.They proposed a language model that gives a probability to each word in the citation context sentences. They then score the importance of sentences in the original paper.

.

4.3. EMAIL SUMMARIZATION

When it comes to email summarization, the texts become a bit different. In order to find a context, a whole chain mail threads are to be processed to determine the story. In (Nenkova & Bagga, 2003), they discuss a few methods of how the conversational nature of the text can be used to gather context. Their methods does provide a conclusive evidence as to how useful this method could be and this further, this method could be enhanced by using some visualization techniques as well. While in (Rambow, Shrestha, Chen, & Lauridsen, 2004), they take a bit of a different approach where they determine important features for summarizing the email, where each feature could be one thread of several threads of email. (Newman & Blitzer, 2003) discusses a whole new approach for summarizing an email. They leverage the clustering algorithm to solve the problem of email summarization. In their paper they talk about few clustering algorithms which help them see all the threads of the email as a whole and post application of the mentioned algorithms, an overview is formed.

Below, Table 2. cite the journals or the references from which some techniques were analyzed and furthermore, their benefits or defects are given. Table 2. Illustrates the various methods which were explained in the articles mentioned in Table 1

Table 1

Journal/Conference	Sub Topic	Years (xxxx-xxyy)	# Articles	#Articles Techniques Benefits	#Articles Techniques Drawbacks
2017 International Conference on Computing Methodologies and Communication (ICCMC)	Text Summarization	2017	Automatic text summarization by local scoring and ranking for improving coherence	It has made use of sentence feature metrics to score sentences like "sentence to sentence cohesion", "word frequency",	it makes use of metrics which might ignore valuable information and thereby making the summary less meaningful. For example, metric like "length of sentence" is used to avoid selecting too short or too long sentences of the document. Doing this, at times, the model might overlook some information which might have been present in those sentences but were not taken into account while summarizing the text.

2017 International Conference on Big Data, IoT and Data Science (BID)	Text Summarization	2017	Automatic text summarization of news articles	The lexical chain generation proposed by Silber and McCoy algorithm has linear run time complexity. Further, certain issues were resolved in both algorithms by implementing pronoun resolution and enhanced sentence scoring to leverage the structure of news articles.	One of the lexical chain generation algorithm adopted was proposed by (Barzilay & Elhadad, 2000) has exponential run time complexity.
Artificial Intelligence Review archive Volume 47 Issue 1, January 2017 Pages 1-66	Text Summarization	2017	Recent automatic text summarization techniques: a survey	This paper talks about a variety of techniques which have their own benefits. 1. Trained Summarizer and Latent Sematic Analysis – uses a modified corpus-based approach and a TRM technique based on latent semantic analysis. The summarizer is based on a function that assesses main sentences/words for things like location, keyword, likeness to title, and centrality in order to generate summaries. The score function is optimized using stochastic techniques such as genetic algorithms. 2. Information retrieval performance has greatly improved	There are some drawbacks of the approaches used which are as follows:Trained Summarizer and Latent Sematic analysis – the summaries generated are not very consistent with topic and the sentences don't correlate so much at times. Feature weights of score function produced by GA fail to consistently give the best results for the test corpus. Obtaining the appropriate dimension reduction ratio and explaining LSA effects are tough in the LSA+TRM technique. Moreover, the time complexity to compute SVD is quite high. 2. Using a sentence-based abstraction technique to extract data – In this approach, only casual coherence is considered whereas

				because of the use of a sentence-based abstraction technique. Sentences that represent the central notion are linked. 3. Understanding and summarizing documents using a document concept lattice – In comparison to existing sentence grouping and sentence scoring algorithms, the suggested approach performs exceptionally well. 4. Sentence extraction using text summarization based on context and statistics – This	space and time which are inter-related links which makes sense out of the document as a whole are also required for representing behavioral context. 3. Through the use of a document concept lattice, it is possible to comprehend and summarize text – time complexity for generating a DCI is high because it considers all possible combinations. 4. Sentence extraction through contextual and statistical based summarization text.

				method outperforms other methods when it comes to summarizing single and numerous documents. 5. Linguistic consistency and subjective opinions are used to summarize e-mails – It has a superior runtime performance and a high accuracy score, according to the evaluation results. Furthermore, the strategy is more accurate than the Page Ranking algorithm. 6. Automatic production of generic document summaries using non-negative matrix factorization — Performance evaluation using t-test has demonstrated that the hypothesis is almost completely followed, although there are a few problems. 7. MR, GA, FFNN, GMM, and PNN models for automatic text summarization – Because there isn't a lot of data on religious and political items, the feature bushy path produces the best results, while the feature presence of numbers produces the worst results. 8. Applying regression models to query-based summarization of numerous documents – The results show that regression models outperform learning to rank and classification models when it comes to calculating the relevance of sentences. 9.	

				This approach, MCMR with B&B algorithm, surpasses all other systems in terms of text summarization coverage and redundancy. It demonstrates that the outcomes are dependent on how comparable the documents are. 10. GRAPHSUM, a graph-based summarizer that explores correlations between numerous terms, outperforms a huge number of state-of-the-art approaches, some of which rely heavily on sophisticated semantic-based models or complex language models.	

International Journal of Advanced Computer Science and Applications	Text Summarization	2017	Text Summarization Techniques: A Brief Survey	This paper talks about automatic evaluation metrics which help in closely evaluating a summary generated by a model comparing to a summary	In Latent Semantic Analysis, In (Gong & Liu, 2001) method chose one sentence for each topic. Therefore, it kept the number of topics as a whole. The drawback for this approach was that a topic can require more than a sentence to exhibit the idea.

For the above table, a file of articles or in the last 20 years were thoroughly analyzed, pertaining to the different techniques of text summarization. These journals have different techniques which has been mentioned above with their benefits and drawbacks.

Therefore, these models can be used as an instrument to analyze the way different text summarization techniques were implemented. In fact, they are powerful tools that can help us in making informed decisions as to what is the advantages and disadvantages of each techniques. Moreover, we can come to conclusions as to what scenarios or in which situation they perform better. Also, these models can provide organizations a comparison of their reality against industry standards, supporting them on defining priorities and achieving their business goals.

After the selection, a deep analysis of each model was conducted using the information provided on the previous mentioned studies, as well as specific information collected and analyzed from each selected model. The results of the analysis are presented in below table.

Table 2

Subject/Model	Year	Author(s)	Reference(s)	Brief Description
Text summary using a trained summarizer and latent semantic analysis	2005	Jen-Yuan Yeh, Hao-Ren Ke, Wei-Pang Yang, I-Heng Meng	(Yeh, Ke, Yang, & Meng, 2005)	Extractive summarization is the technique where relevant sentences are selected to compose the summary of the given text. Before forming the summary, the ratio factor is taken as input which decides how big or small the summary is going to be. The approach mainly scores the sentences based on functions/metrics which determine whether that sentence would be selected in the composition of the summary. This section explains the latest text summarization approaches used in the last decade employing a sentence's semantic representation. LSA is used to extract latent structures from a document. MCBA and LSA+TRM approach mainly summarizes a single document and compose extract-based summaries. Conclusion: Cen and R2T are the two important features and mix of features like Pos, +ve keyword, Cen and R2T are the best. GA gives us a combination of feature weights in the training phase. LSA+TRM performs better than keyword-based text summarization techniques in both single-document as well as multi-document.
Information extraction using sentence-based abstraction technique	2006	Samuel W.K.Chan	(W.K.Chan, 2006)	(W.K.Chan, 2006) created a new quantifiable approach for creating summaries that takes phrases from the text's most relevant portion. This method employs a shallow linguistic extraction strategy. This method uses a sentence-based abstraction technique to extract information. A discourse network is built to represent speech that contains not just sentence boundaries but also text consisting of interconnected components as a single unit rather than discrete sentences in a series. The smallest unit of interaction in a discourse network is the discourse segment. Textual continuity is used in this approach to connect the segments via a discourse network. The two quantitative coefficients used to assess the degree of discourse continuity are cohesion and coherence. Cohesion is the representation of connection between sentences in close segments, and it is conveyed in a text by practical and syntactic relations between sentences and clauses. Referential cohesion, lexical cohesion, and verb cohesion are some of the cohesion variables considered. The link between neighboring segments that is not obvious in the text is referred to as coherence.

Text understanding and summarization through document concept lattice	2006	Shiren Ye, Tat-Seng Chua, Min-Yen Kan, Long Qiu	(Ye, Chua, Kan, & Qiu, 2007)	(Ye, Chua, Kan, & Qiu, 2007) suggested the Document Concept Lattice (DCL), a data structure in which the source document's concepts are encoded by a direct acyclic graph, with nodes representing the set of overlapping concepts. Concepts are words that reflect concrete entities and their behaviors in this context. As a result, concepts serve as indicators of key facts and as a tool for answering critical issues. The summarization method uses DCL to find a globally optimal collection of sentences that reflect the greatest number of conceivable concepts with the fewest number of words. This is performed using the representational power of a summary, which is a fitness metric for the summary. Dynamic programming is used to explore the search space of DCL in three steps: (a) A group of important internal nodes is chosen, (b) sentences with the highest representative power are chosen from these key internal nodes, and (c) after analyzing an amount and variety of the chosen sentences, the best combination that results in the least amount of answer loss is chosen. Finally, this method generates an output summary containing the collection of statements having the greatest representational power. Conclusion: When compared to existing sentence grouping and sentence scoring algorithms, the suggested methodology turns out to be competitive.
Sentence extraction through contextual information and statistical based summarization of text	2009	Youngjoong Ko, Jungyun Seo	(Ko & Seo, 2008)	(Ko & Seo, 2008) suggested an excellent method for text summarizing that uses contextual information and statistical methodologies to extract important sentences. Using a sliding window mechanism, two successive phrases are first concatenated to generate a Bi-Gram Pseudo Sentence (BGPS) (Ko & Seo, Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique, 2004) Because BGPS has more features (words) than a single sentence, it overcomes the problem of feature sparsity caused by extracting features from a single sentence. The suggested technique handles two types of sentence extraction tasks. Many relevant BGPS are identified from the target document in the first stage. After that, each BGPS is broken into two separate sentences. The separated sentences are worked on in the second stage, and essential sentences are extracted to provide a final summary. The title technique, aggregation similarity method, location method, frequency method, and tf-based query method are the hybrid statistical sentence extraction approaches employed here. The suggested method is also used in multi-document summarizing, where two sentence extraction procedures are used to

				provide a summary for each document in the document cluster via the primary process of sentence extraction. The resultant summary of the document cluster is then generated via secondary process using the summaries collected in the main process. *Conclusion: This method outperforms other methods when it comes to summarizing single and multiple documents.*
Email summaries based on conversational cohesiveness and subjective opinions	2008	Giuseppe Carenini, Raymond T. Ng, Xiaodong Zhou	Carenini et al. 2008	New ways for summarizing email exchanges were proposed by (Carenini, Ng, & Zhou, 2008). Initially, a segment quotation network is constructed based on a dialogue including a few emails, with nodes representing discrete fragments and edges representing fragment answering relationships. Then, using this fragment quotation graph, create a sentence quotation graph in which each sentence in the email exchange is represented by a distinct node in the graph, and a replying relationship is represented by an edge connecting two nodes. Three types of cohesion metrics are investigated in order to apply weights to the edges: clue words (stem-dependent), semantic similarity (WordNet-dependent), and cosine similarity (TF-IDF dependent). The subject of extractive summarization is thought to be a node ranking problem. To compute each sentence's score (node), the Generalized Clue Word Summarizer (CWS) and Page-Rank, i.e., the two graph-based summarization algorithms, are utilized, and then highly scored sentences are used to construct the summary. The overall weight of all outbound and inbound edges of a node is summed to calculate the grade of a sentence in Generalized Clue Word Summarizer, but it does not account the node's relevance (sentence). The weights of outgoing and incoming edges, as well as the relevance of nodes, are taken into account by a Page-Rank-based summarizer. To present a summarizing methodology that helps select more essential sentences, subjective opinions are combined with graph-based approaches. Subjective judgments are combined with the best cohesiveness metric to achieve better results. The sentence with the most subjective words is regarded as a crucial sentence for the summary. The two items of subjective words and phrases evaluated in this technique are OpFind and Opbear.
Automatic creation of generic	2009	Lee J-H	Lee et al 2009	Through Non-negative Matrix Factorization (NMF), (Lee, Park, Ahn, & Kim, 2009) proposed an unsupervised summarizing strategy for general-purpose materials . Singular vectors are employed in the Latent Semantic Analysis (LSA) approach for sentence selection, and they might have negative values and are not scarce, therefore

document summaries through non - negative matrix factorization				this approach could not naturally catch the meaning of semantic features that are highly sparse and have a limited view of meaning. As a result, summarization systems based on LSA are unable to choose meaningful phrases. As a result, elements of semantic feature vectors in the suggested method exclusively contain non-negative values and are also extremely sparse, allowing semantic characteristics to be easily read. A sentence can be represented by a linear combination of certain significant semantic elements. As a result, subtopics in a document can be quickly identified, and there's a better probability of extracting relevant lines. A method for picking phrases to construct general document summaries is suggested using NMF, in which a content is first pre-processed and then summarized. To generate a non-negative semantic feature matrix, NMF is used to a term-by-sentence matrix. For each sentence, generic relevance is calculated, which indicates how much a sentence explains.
Query based summarization of multiple documents by applying regression models	2011	Ouyang Y	Ouyang, Li, Li, & Lu, 2011	(Ouyang, Li, Li, & Lu, 2011) suggested a method for ranking phrases in query-based summarization of numerous manuscripts using regression models. Three query-dependent features, such as named-entity matching, word-matching, and semantic matching, and four query-independent features, such as sentence position, named entity, word TF-IDF, and stop-word penalty, are used in this methodology to choose main sentences in query-based summarization of multiple documents. To begin with, human summaries generate "false" training data. Then, using different methods based on the N-gram methodology that calculate "nearly true" relevance ratings of phrases are created and analyzed using this training data and their collection of texts, and a mapping function is learned using this training data via a collection of previously specified features of sentences. Then, using this learned function, the significance of sentences in the test data is estimated. An efficient data collection of training data for learning regression models requires two things: (a) an appropriate group of topics with correctly handwritten summaries, and (b) a good approach for computing the relevance of words. The Maximal Marginal Relevance (MMR) technique is used to remove redundancy from the summary.
Automatic text summarization using MR, GA,	2009	Mohamed Abdel Fattah, Fuji Ren	(Fattah and Ren 2009)	With the use of a few statistical features, (Fattah & Ren, 2009) suggested an approach to improve content selection in automatic text summarization. As a trainable summarizer, this method relies on distinct statistical aspects in each sentence to generate summaries. Position of Sentence (Pos), +ve keyword, -ve keyword, +ve

FFNN, GMM and PNN based models				keyword, +ve keyword, +ve keyword, +ve keyword, +ve keyword, +ve keyword, +ve keyword, +ve keyword, +ve keyword, R2T, Centrality of Sentence (Cen), Presence of Name Entity in Sentence (PNE), Presence of Numbers in Sentence (PN), Bushy Path of Sentence (BP), Relative Length of Sentence (RL), and Aggregate Similarity (AS) are all measures of sentence similarity. Genetic Algorithm (GA) and Mathematical Regression (MR) models have been trained to acquire an optimal mix of feature weights by mixing all of these features. For sentence categorization, feed forward neural networks (FFNN) and probabilistic neural networks (PNN) are utilized. Some text features, such as the +ve and -ve keywords, are language-dependent, while eight others are not. All of the above-mentioned variables are taken into account when calculating a sentence's weighted score function. All sentences in a document are ranked in decreasing order of their scores, and a highly scored cluster of sentences is utilized to generate a summary of the content using various compression rates (10, 20, 30 percent used here). Conclusion: The results demonstrate that feature BP is the most essential text feature since it produces the best results, while feature PND produces the worst results because statistical data is absent from religious and political pieces. Because it could model arbitrary densities, the GMM approach produced the best results of all the strategies.
Maximum coverage and minimum redundancy in summarization of text	2011	Alguliev, R. M	Alguliev, Aliguliyev, Hajirahimova, & Mehdiyev, 2011	(Alguliev, Aliguliyev, Hajirahimova, & Mehdiyev, 2011) introduced an unsupervised summarizing model for generic text as an Integer Linear Programming problem (ILP) that immediately detects essential sentences from the article as well as the full article's relevant information. Maximum Coverage and Minimum Redundancy is the name of this strategy (MCMR). This method aims to improve three key aspects of a summary: (a) relevance, (b) redundancy, and (c) length. A subset of sentences from the document collection's relevant text is picked. Then, using NGD-based similarity (Normalized Google Distance) and cosine similarity, similarity between the summary and the document collection is computed, and this similarity must be maximized. An objective function is developed and must be maximized to ensure that the summary contains the important content found in the document collection and that the summary does not contain a significant number of phrases that communicate the same information. At the same time, the length of the summary must be limited. Lastly, an empirical function is created by linearly combining the cosine similarity-

				based and NGD-based similarity empirical functions, and this combined empirical function must be maximized. This technique to summarizing is incorporated as an optimization problem that aims to find a global solution to the problem. The Branch & Bound algorithm (B&B) and the Binary Swarm Optimization method are the algorithms used to address the ILP problem. Conclusion: This method, which combines MCMR with the B&B algorithm, surpasses all others. It demonstrates that summarizing outcomes are dependent on similarity measurements. It is also proved through tests that using cosine similarity and NGD-based similarity metrics together produces better results than using them separately.
Summarization of documents through a progressive technique for selection of sentences	2013	Ouyang Y	Ouyang, Li, Zhang, Li, & Lu, 2013	(Ouyang, Li, Zhang, Li, & Lu, 2013) proposed a new progressive technique for generating a summary based on the selection of "novel and salient" sentences. Subsuming relationship between two sentences, i.e., an irregular relationship between sentences that shows the level of recommendation of one phrase by another. In order to ascertain the link between two phrases, the relationship between their concepts must be found. The association between concepts is then discovered by using a coverage-based measure to discover the relationship between words. A Direct Acyclic Graph (DAG) is used to organize all of the words that appear in the found word relations. A progressive strategy for sentence selection is created on the basis of an asymmetric relationship between sentences, in which a sentence is either picked as a novel general statement or as a supporting sentence. The following two methods are used to choose new and relevant sentences in this method: (a) discovered concepts are only included during the assessment of sentence relevance to assure sentence originality, and (b) for now, the relationship between sentences is used to improve the saliency measure. To implement this strategy, a random walk on the DAG from the central node to its nearby nodes is performed, with the goal of covering the central words first and then reaching the greatest amount of words via word relations. Redundancy is eliminated by punishing repetitive words, resulting in fresh concepts being introduced each time a new phrase is chosen. Conclusion: In terms of generating summaries with improved saliency and coverage, the Progressive system surpasses the traditional Sequential approach.
Evaluation of	2013	Ferreira, Rafael; Cabral, Luciano de	Ferreira, et al.,	In the recent decade, (Ferreira, et al., 2013) incorporated fifteen scoring techniques that had been referenced in the research. ROUGE (Lin 2004) is used for quantitative

sentence scoring methods for extractive summarization of text		Souza; Lins, Rafael Dueire; Silva, Gabriel Pereira e; Freitas, Fred; Cavalcanti, George D.C.; Lima, Rinaldo; Simske, Steven J.; Favaro, Luciano	20132013	evaluation, while the number of sentences that are similar among the machine-generated and human-made summaries is counted for qualitative evaluation. The processing time of each algorithm is taken into account. Word scoring, sentence scoring, and graph scoring approaches are used to pick relevant sentences. The most essential terms are given scores in the word scoring approach. Word frequency, TF/IDF, upper case, proper noun, word co-occurrence, and lexical similarity are among the approaches used to score words. The properties of sentences are examined in the sentence scoring approach. The existence of cues, numerical data, sentence length, sentence position, and sentence centrality are all factors in sentence scoring. Scores are determined using the graph scoring approach by looking at the relationships between sentences. Text rank, bushy path of the node, and aggregate similarity are all graph scoring approaches. The six common concerns of stop words, structural transformation, comparable semantics, ambiguity, redundancy, and co-reference are then explored, along with some suggestions for advancing sentence score outcomes.
Exploring correlations among multiple terms through a graph-based summarizer, GRAPHSUM	2013	Baralis, Elena; Cagliero, Luca; Mahoto, Naeem; Fiori, Alessandro	Baralis, Cagliero, Mahoto, & Fiori, 2013	GRAPHSUM, a new graph-based, general-purpose summarizer for summarizing numerous documents, was proposed by (Baralis, Cagliero, Mahoto, & Fiori, 2013). This method investigates and applies association rules, a data mining methodology for finding connections between several terms. It is not reliant on sophisticated semantic models (like taxonomies or ontologies). The document collection is organized as a transactional dataset after preprocessing so that association rule mining may be conducted on it. Then, from the transactional dataset, frequently recurring itemsets with high correlations among the terms are identified, and a correlation graph is constructed from these terms, which will aid in the selection of significant lines for the summary. The Apriori algorithm is used to mine frequently recurring itemsets, and the support measure is employed for this job. The lift measure indicates the intensity of relationship between two terms and is used to evaluate positive or negative connections between commonly used words. A variation of the classic PageRank graph ranking algorithm is used to determine the relevance of the graph nodes. The graph nodes that have a significant number of positive correlations are placed first, while those that have a negative connection with the adjacent nodes are penalized. For summary creation, the sentences that are the most appropriate for the correlation

				graph and have a high relevance score are picked. The greedy algorithm is employed to select sentences in this case. GRAPHSUM performs better over a wide range of state-of-the-art techniques, some of which rely heavily on highly developed semantic-based models or complicated language processes.
Incorporating various levels of language analysis for tackling redundancy in text summarization	2013	Elena Lloret, Manuel Palomar	Lloret & Palomar, 2013	(Lloret & Palomar, 2013) provided a method for detecting redundant information based on three layers of language analysis: lexical, syntactic, and semantic. Cosine similarity is utilized in the lexical based technique to detect similarity between sentences in two sources. Those sentences that have a cosine similarity greater than a certain threshold are considered repetitive, and they are all eliminated. In a syntactic-based method, entailment relations are computed between pairs of phrases to determine whether the meaning of one sentence can be deduced from the meaning of the other sentence. If a positive entailment is obtained, the second sentence is deemed superfluous and eliminated. Sentence alignment is determined at the document level between a set of linked documents using a open source available Champollian Tool Kit in a semantic-based manner. Syntactic and semantic techniques are preferable than lexical approaches that rely on cosine similarity. Text summarization can be done in two ways. Before the material is summarized, unnecessary sentences are deleted in the first technique. The set of useful sentences is then given to the summarization system, which uses statistical (term frequency) and linguistic (code quantity principle) factors to select essential sentences, as well as a summary.

5. METHODOLOGY

This study will be conducted through a qualitative research, based on a well-structured parameter for comparing the different techniques adopted for text summarization techniques. This is the basis to compare different techniques and shed light on which techniques would be more useful in (if any) particular situations. What are their drawbacks and advantages?

5.1. DESIGN SEARCH RESEARCH

Design Science Research is a form of investigation that entails building or improving something in a novel way in response to a specific challenge.

The quest for a solution based on extensive scientific investigation ensures that the final proposed artifact is coherent and credible. A crucial phase that should not be overlooked is good communication of the finished product (Hevner, March, Park, & Ram, 2004).
Each of the six key stages of DSR methodology, as shown in Figure 5, will be discussed in greater detail right away.

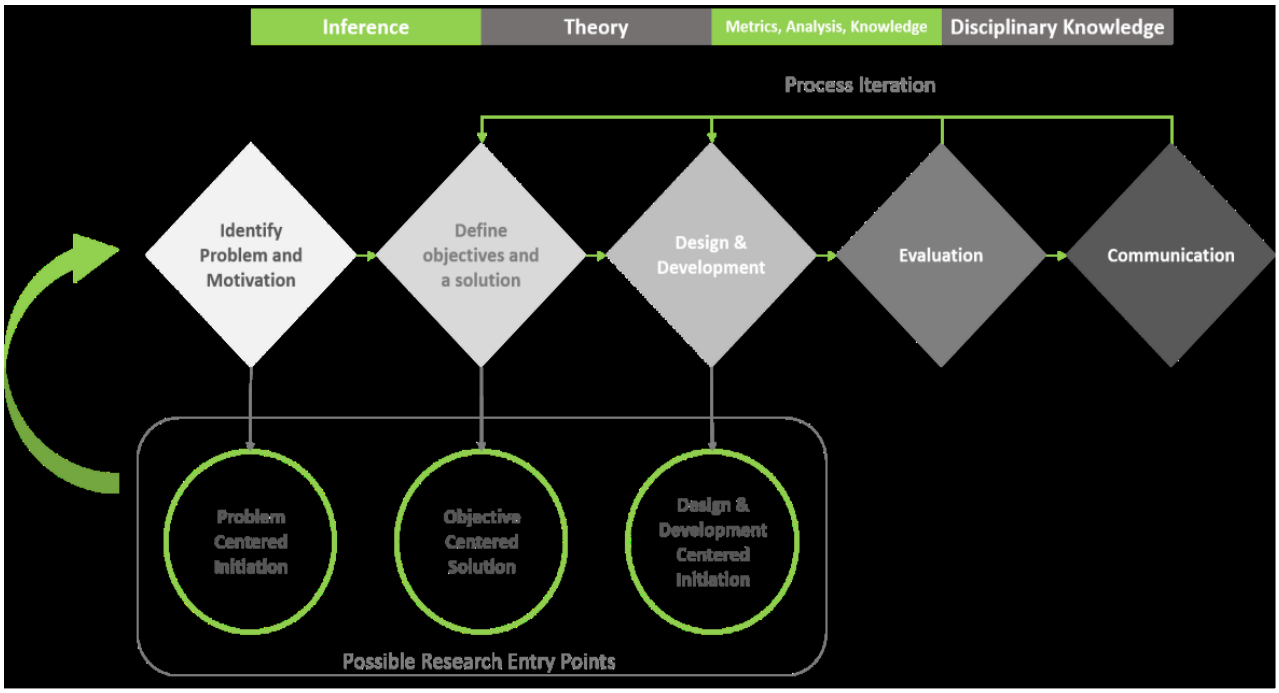


Figure 3. DSR Method Adaptation (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007)

Identify problem and motivation

Define the research challenge in detail and justify the importance of a solution.
Begin by establishing a testable theory that leads to a research problem by demonstrating to stakeholders the value of an effective solution and what they will gain from its result (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Define objectives and a solution

Clearly define goals (quantitative or qualitative) to establish the foundation for a solution based on the problem characterization and what can and cannot be done (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Design and Development

The goal of the design and development stages is to create knowledge through the design and development of the artifact itself (Gregor & Hevner, 2013). This could be accomplished by breaking down the major scientific problem into smaller components (Hevner, March, Park, & Ram, 2004). To have an effective/ clear structure in the next phase, it is necessary to have a clear grasp of the solution value and to defend it with some theoretical foundation (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). A solution that must meet business requirements (Hevner, March, Park, & Ram, 2004).

To gain the appropriate theoretical basis, it is essential to do research and collect knowledge about the present status of the problem and existing solutions, as well as to analyze direct and indirect solutions and their efficacy (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). With the knowledge, it is possible to develop a solution to meet research and, as a result, business objectives, as well as to debate the usefulness of the suggested artifact (Hevner, March, Park, & Ram, 2004).

Evaluation

To certify an artifact's efficacy, it must be put to use or presented to stakeholders (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007), which must be supported by a clear specification of evaluation methodologies that are suitable for the situation at hand and are based on industry requirements. Because the majority of claims on the final solution are related to performance issues, alignment with business needs is critical (Hevner, March, Park, & Ram, 2004).
Comparing what falls under the purview of the master's thesis with what could be observed in its practical implementation is one technique to evaluate how the answer matches the initial challenge (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).
Although it is critical to emphasize that the primary goal is to "identify how well an artifact works" rather than "theorize or prove anything about why the artifact works" (Hevner, March, Park, & Ram, 2004).
At the end of this phase, it should be determined whether the artifact is ready to be shared with the rest of the world, or whether more effort should be spent improving it to make it more effective/aligned with the original problems (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007).

Communication

While releasing the final artifact to the public is a step in the right direction, it's also critical to let people know how unique and successful the artifact is in solving the highlighted problems (Peffers, Tuunanen, Rothenberger, & Chatterjee, 2007). It is critical to discuss how the artifact was created and the review process that led to its validation throughout this communication (Hevner, March, Park, & Ram, 2004).

It should be conveyed to technical and management audiences in order to gather input to enhance the solution, both in terms of business and technology, for future implementations (Hevner, March, Park, & Ram, 2004).

5.2. STRATEGY

Problem

There are numerous text summarizing approaches, each with its own set of benefits and drawbacks. Some are more computationally complex than others, while others have only been implemented in specific languages. Some utilize more statistical measures to quantitatively address summarizing problems, while others more extractive in nature. Despite all of these possibilities, there is no single approach or methodology that can be used on any type of text. We need to know which strategy or technique to use in various situations.

Objective

After stating the topic, our goal in this paper will be to research and assess several strategies, as well as to describe their benefits and drawbacks, as well as the situations and circumstances in which they might be employed. In the same case, not all methods would perform the same. As a result, we would do our best to produce a fair comparison and highlight the techniques' or methodology' limitations.

Design and Development

Initially, a number of research publications on text summarization approaches were examined. Some of the strategies for examining its algorithm, time complexity, the data it was implemented on, how efficient the algorithm is, how useful the generated summary is, and whether it was an abstractive or extractive based methodology have been detailed in depth above.

6. PROPOSAL OF A FRAMEWORK ON SCENARIOS OF TEXT SUMMARIZATION TECHNIQUES

6.1. PROPOSAL

Although text summarization has a vast number of techniques to offer, it was not possible to cover all of those here and a such only few were selected, which were studied here aforementioned in the above tables. The following table below compares those above techniques in terms of accuracy and time complexity, applicability. Although this table does not give a fair comparison since, all these techniques were not applied on the same document and for the same situations.

Table 3

Techniques	Parameters			
	Accuracy	Speed	Applicability on different language	Scenarios applicable
The lexical chain generation	Accuracy is better	Has linear run time complexity	For example, Bengali	Although this method can be applied to multiple situations, most research papers state its main applicability in World Wide Web.
Latent Semantic Analysis	Certain combinations show different accuracy mentioned below	Linear Time complexity.	For example, Bengali, Hindi	LSA now scales to ca. 100 million-word corpora by larger computer memory and new algorithms.
Query based summarization of multiple documents by	Results demonstrate that for computing the importance of In	The speed varies with documents explained in detail	Although, any paper related to this technique has not	summarizing research papers of a specific domain, biomedical documents for better accuracy

applying regression models	comparison to training to score and classifying models, regression models perform better.	below	yet been applied to other language. But this technique should not have any issues (technical) if applied to other language.	
In summarizing text, maximum scope and desired minimal repetition	Accuracy is 97% according to (HoudaOufaida, OmarNouali, & PhilippeBlache, 2014). Although this is just one sample.	Computational time is proportional to O(X*Y) where X and Y are different terms in the distance matrix used to discern the similarity.	Tested in languages like Arabic, Czech, English, French, Greek, Hebrew and Hindi	single- and multi-document summarization. In both tasks, documents are split into sentences in preprocessing
Evolutionary optimization algorithm for summarizing multiple documents	Accuracy is usually good if the algorithm is run making sure that the whole search space is covered and not stuck at local maxima	Time complexity for these algorithms is usually pretty high as it has to make sure that the whole search space is covered during the run-time.	This proposed method has not yet been applied in other languages.	Digital archives of governmental documents

The lexical chain generation - Word Sense Disambiguation (WSD) accuracy is better. The algorithm proposed by Silber and McCoy has linear run time complexity. Tested in different languages apart from English. For example, Bengali. Although this method can be applied to multiple situations,

most research papers state its main applicability in World Wide Web. The precision and recall of this method is given below with different text files on the X axis in the below chart which gives an idea of how it performs.

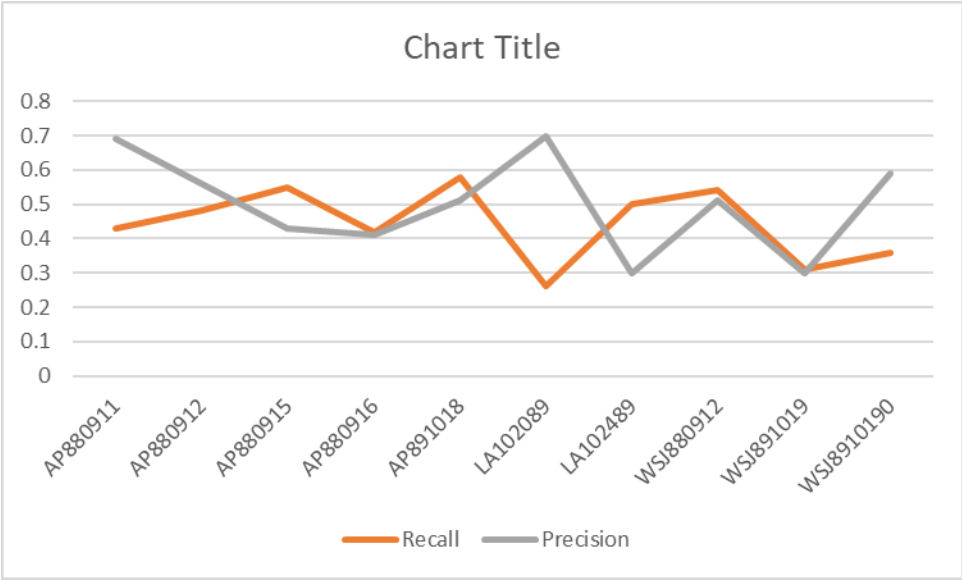


Figure 8 : Precision and Recall of different text files

Latent Semantic Analysis - Certain combinations of parameter show a lack of precision and can even lead to deceptive similarity measurements For tiny datasets, choices that respect the fundamental data's potential richness produce superior results: either no SVD or an SVD with a large number

of dimensions. Furthermore, under certain circumstances, weighing has a positive effect. A universal unified indexing and weighting (and SVD, if used) method does not produce worse results than an individual, case-based indexing and weighting approach for a group of small datasets. For each term *tj*, the time cost for computing partial similarities between *di* and *tj* for all $di \in D$ is derived as $O(N)$. Since only the partial similarities bigger than θ are considering for creating index nodes, so the total time cost for creating $\langle di, \text{PartialSim}\theta(di, tj)\rangle$ in $I\theta(tj)$ for all $di \in D$ is derived as $O(\varepsilon tN)$, where εt is ratio of the partial similarities lower than θ between term *tj* and all document $di \in D$. And then the total time cost for computing partial similarities and creating index nodes is derived as $O((1+\epsilon)N)$. And finally, the time cost of this algorithm is derived as $O(1+(1+\epsilon)rN)$, where ϵ is average εt for all $tj \in \{tj \mid\mid VtT(tj:) \neq 0\}$. The time cost of this algorithm is determined by the size of matrix VtT and threshold θ. Tested in different languages apart from English. For example, Bengali, Hindi. LSA now scales to ca. 100 million-word corpora by larger computer memory and new algorithms

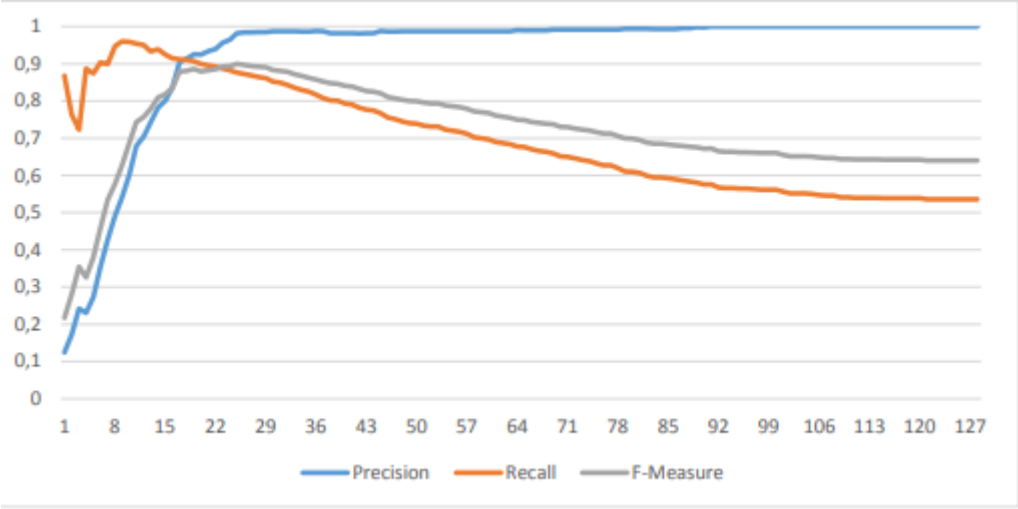


Figure 9 : Precision, Recall and F-Measure for different values of k applying LSA

Application of regression models to query-based summary of various documents - The findings show that regression models perform better learning to rank and classification models when it comes to calculating the significance of phrases. The number of nodes and edges visited by A* search, reflecting the space and time complexity of the algorithm, as a function of the number of sentences in the document set being summarized, all three

heuristics show an empirical increase in complexity that is roughly linear in the document size, although there are some notable outliers, particularly for the uniform heuristic. Although, any paper related to this technique has not yet been applied to other language. But this technique should not have any issues (technical) if applied to other language. summarizing research papers of a specific domain, biomedical documents for better accuracy. These models were run on DUC 2005 data set. Table 4 below provides the average ROUGE-2 and ROUGE-SU4 scores and the corresponding 95% confidential intervals. As expected, regression models outperform both classification models and ranking models.

Table 4

Model	Average ROUGE-2 (CI)	Average ROUGE-SU4 (CI)
Regression	0.0757	0.1335
Ranking	0.0715	0.1287
Classification	0.0641	0.1208

Ref - DOI: 10.1016/j.ipm.2010.03.005

In summarizing text, maximum coverage and desired minimal repetition - According to the authors, the results were very satisfying, sometimes reaching a 97% accuracy threshold. Computational time is not negligible, although more new algorithms are being explored. It has been tested in languages like Arabic, Czech, English, French, Greek, Hebrew and Hindi. single- and multi-document summarization. In both tasks, documents are split into sentences in preprocessing. We observe that the result of this method directly depends on the optimization algorithm. As shown in Table 5, among two algorithms B&B and PSO, the best result is obtained by the B&B. It is observed that this method MCMR (B&B) with the B&B optimization algorithm demonstrates the best ROUGE values and outperforms all the other systems.

Table 5

Methods	Improvement of the method MCMR (B&B) (%)	
	ROUGE-2	ROUGE-SU4
MCMR (B&B)	0.00	0.00
MCMR (PSO)	4.81	3.30
PNR^2	36.42	35.79
PPRSum	2.18	2.51
GSPSum	10.00	7.02
AdaSum	4.18	3.61

Ref: DOI: 10.1016/j.eswa.2011.05.033

Evolutionary optimization algorithm for summarizing multiple documents - The proposed method leads to competitive performance. Statistical results depict that this method performs better than other baseline methods. Since this method has a broad search space, the time complexity for reaching an optimal accuracy comes with the price of time. This proposed method has not yet been applied in other languages. Digital archives of governmental documents. Among the methods, in general WFS-NMF achieves the highest ROUGE-2, ROUGE-L, and ROUGE-SU scores. This observation demonstrates that the sentence feature selection is effective and the weights on document side help the sentence weighting process. The figure 6 below shows the different methods incorporated on DUC2002 and DUC2004 documents with their ROUGE scores on the Y axis.

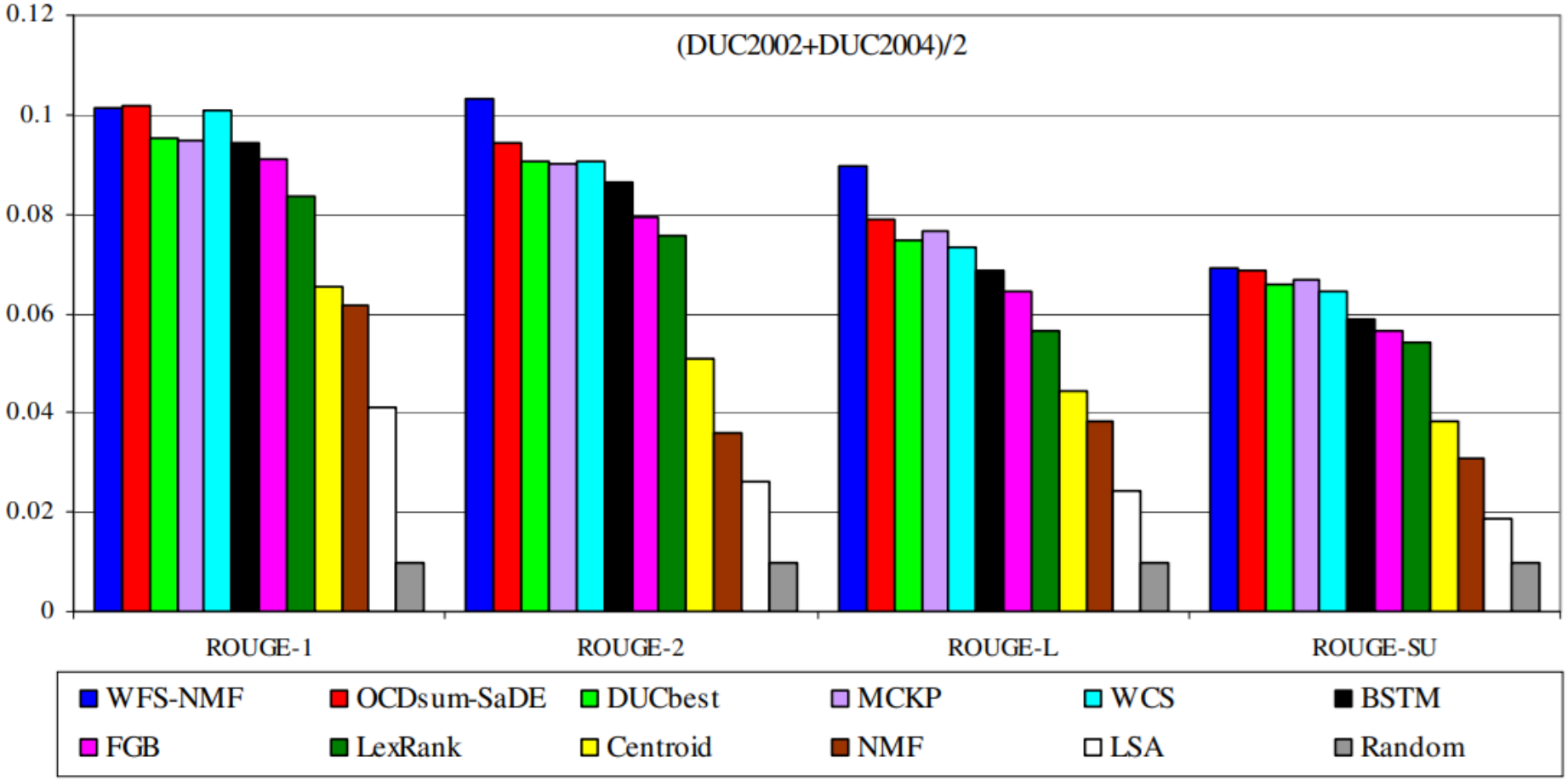


Figure 10 : Overall Comparison of the methods

R.M. Alguliev et al. / Expert Systems with Applications 40 (2013) 1675–1689 - https://doi.org/10.1016/j.eswa.2012.09.014

6.2. VALIDATION

I tried my best to try to compare the different approaches and I hope this helps companies/research to have a better grasp of the advantages and disadvantages.

While validating this section, I had few questions which needed to be answered like:

- Why the study should be done,

  There is a myriad of approaches used in NLP which needs to be studied and analyzed and their advantages and disadvantages documented which saves a lot of time going through all the trial-and-error methods before coming to a conclusion as to which model fits one's situation well and would solve specific problems. But none of them were documented in one particular paper, thereby this paper helps troubleshoot those kinds of problem for beginner in NLP field.

- The potential implications emerging from your proposed study of the research problem, and

  These studies and analysis are documented to help the concerned. And in doing so, I come across the shortfalls and the benefits of using one technique. And the shortfalls or the benefits are just not algorithm based but depends on the context the problem is being used on. So, it was not totally fair to compare these algorithms based on their accuracies having different contextual problems.

- A sense of how your study fits within the broader scholarship about the research problem.

  The study is an eye opener for those new to the field of NLP and it definitely helps industries and researchers solve a specific problem without having to go through all the methods which would definitely save lot of time.

7. CONCLUSIONS

As the Internet has grown in popularity, a vast amount of information has become available. Summarizing vast amounts of text is challenging for humans. In this age of information overload, automatic summarizing technologies are in high demand.

Various extraction methodologies for single and multi-document summarization were highlighted in this research. Topic representation approaches, frequency-driven methods, graph-based and machine learning techniques were described as some of the most often utilized methodologies. Although it is impossible to elucidate all of the many methods and approaches in my thesis, it does provide a good overview of recent trends and advancements in automatic summarizing methods and describes the current state-of-the-art in this field.

Limitations

One of the main limitations of this report is that it wasn't validated by lot of people given the fewer number of experts in this field. With that goes the unsaid, that this paper doesn't document all the NLP techniques, which is quite a broad field.

8. REFERENCES

A., N., & K., M. (2012). A Survey of Text Summarization Techniques. Em A. C., & Z. C., *Mining Text Data* (pp. 43-76). Boston, MA: Springer, Boston, MA.

Alguliev, R. M., Aliguliyev, R. M., Hajirahimova, M. S., & Mehdiyev, C. A. (2011). MCMR: Maximum coverage and minimum redundant text summarization model. *Expert Systems with Applications*, 14514-14522.

Allahyari, M., Pouriyeh, S., Assefi, M., Safaei, S., Trippe, E. D., Gutierrez, J. B., & Kochut, K. (2017). Text Summarization Techniques: A Brief Survey.

Amitay, E., & Paris, C. (2000). Automatically Summarising Web Sites - Is There A Way Around It? *CIKM00: Proceedings of the ninth international conference on Information and knowledge management* (pp. 173–179). Association for Computing MachineryNew YorkNYUnited States.

Baralis, E., Cagliero, L., Mahoto, N., & Fiori, A. (2013). GraphSum: Discovering correlations among multiple terms for graph-based summarization. Em *Information Sciences* (pp. 96-109).

Barzilay, R., & Elhadad, M. (2000). Using Lexical Chains for Text Summarization.

Carenini, G., Ng, R. T., & Zhou, X. (2008). Summarizing Emails with Conversational Cohesion and Subjectivity. (pp. 353–361). Association for Computational Linguistics.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Fattah, M. A., & Ren, F. (2009). GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech & Language*, 126-144.

Ferreira, R., Cabral, L. d., Lins, R. D., Silva, G. P., Freitas, F., Cavalcanti, G. D., . . . Favaro, L. (2013). Assessing sentence scoring techniques for extractive text summarization. Em *Expert Systems with Applications* (pp. 5755-5764).

Gong, Y., & Liu, X. (2001). Generic text summarization using relevance measure and latent semantic analysis. *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 19-25). SIGIR '01.

HoudaOufaida, OmarNouali, & PhilippeBlache. (2014). Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. *Journal of King Saud University - Computer and Information Sciences*, 450-461.

III, H. D., & Marcu, D. (2006). Bayesian Query-Focused Summarization. *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 305-312). ACL-44.

Jones, K. S. (2004). A statistical interpretation of term specificity. *Journal of Documentation Volume 60 Number 5*, 493-502.

Ko, Y., & Seo, J. (2004). Learning with Unlabeled Data for Text Categorization Using a Bootstrapping and a Feature Projection Technique. *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04)*, (pp. 255–262).

Ko, Y., & Seo, J. (2008). An effective sentence-extraction technique using contextual information and statistical approaches for text summarization. Em *Pattern Recognition Letters* (pp. 1366-1371).

Lee, J.-H., Park, S., Ahn, C.-M., & Kim, D. (2009). Automatic generic document summarization based on non-negative. *Information Processing and Management*, 20-34.

Lloret, E., & Palomar, M. (2013). Tackling redundancy in text summarization through different levels of language analysis. Em *Computer Standards & Interfaces* (pp. 507-518).

Luhn, H. P. (1958). The Automatic Creation of Literature Abstracts. *IBM JOURNAL APRIL 1958*.

Mei, Q., & Zhai, C. (2008). Generating Impact-Based Summaries for Scientific Literature. *Proceedings of ACL-08: HLT* (pp. 816–824). Columbus: Association for Computational Linguistics.

Nenkova, A., & Bagga, A. (2003). Facilitating email thread access by extractive summary generation. *Recent Advances in Natural Language Processing III: Selected papers from RANLP 2003*, pp. 287-.

Newman, P. S., & Blitzer, J. C. (2003). Summarizing Archived Discussions: A Beginning. *Proceedings of the 8th international conference on Intelligent user interfaces* (pp. 273–276). IUI '03.

Ouyang, Y., Li, W., Li, S., & Lu, Q. (2011). Applying regression models to query-focused multi-document summarization. *Information Processing & Management*, 227-237.

Ouyang, Y., Li, W., Zhang, R., Li, S., & Lu, Q. (2013). A progressive sentence selection strategy for document summarization. *Information Processing & Management*, 213-221.

Rambow, O., Shrestha, L., Chen, J., & Lauridsen, C. (2004). Summarizing Email Threads. *Proceedings of HLT-NAACL 2004: Short Papers* (pp. 105–108). HLT-NAACL-Short '04.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Information Processing and Management*.

W.K.Chan, S. (2006). Beyond keyword and cue-phrase matching: A sentence-based abstraction technique for information extraction. Em *Decision Support Systems* (pp. 759-777).

Ye, S., Chua, T.-S., Kan, M.-Y., & Qiu, L. (2007). Document concept lattice for text understanding and summarization. Em *Information Processing & Management* (pp. 1643-1662).

Yeh, J.-Y., Ke, H.-R., Yang, W.-P., & Meng, I.-H. (2005). Text summarization using a trainable summarizer and latent semantic analysis. Em *Information Processing & Management* (pp. 75-95).